A Unified Model for Cross-Domain and Semi-Supervised Named Entity Recognition in Chinese Social Media

Hangfeng He, Xu Sun
MOE Key Laboratory of Computational Linguistics, Peking University
School of Electronics Engineering and Computer Science, Peking University
{hangfenghe, xusun} @pku.edu.cn

Abstract

Named entity recognition (NER) in Chinese social media is important but difficult because of its informality and strong noise. Previous methods only focus on in-domain supervised learning which is limited by the rare annotated data. However, there are enough corpora in formal domains and massive in-domain unannotated texts which can be used to improve the task. We propose a unified model which can learn from out-of-domain corpora and in-domain unannotated texts. The unified model contains two major functions. One is for cross-domain learning and another for semi-supervised learning. Cross-domain learning function can learn out-of-domain information based on domain similarity. Semi-Supervised learning function can learn in-domain unannotated information by self-training. Both learning functions outperform existing methods for NER in Chinese social media. Finally, our unified model yields nearly 11% absolute improvement over previously published results.

Introduction

Named entity recognition (NER) is a basic task in natural language processing (NLP). NER is important and useful for many high-level applications such as information extraction and entity linking. With the development of Internet, more and more researches focus on NER in social media (Li and Liu 2015; Habib and van Keulen 2015; Cherry and Guo 2015; Peng and Dredze 2016b). NER in social media is more challenging because of its informality and strong noise. Although efforts in English have narrowed the gap between social media and formal domains (Cherry and Guo 2015), NER in Chinese social media is still quite hard.

NER is a task to identify names in texts and to assign names with particular types (Sun et al. 2009; Sun 2014; Sun et al. 2014). Previous work for NER in Chinese social media (Peng and Dredze 2015; 2016a) mainly use Condition Random Field (CRF) to deal with the task. They used supervised learning which is limited by rare annotated data. We want to use deep learning methods to learn from out-of-domain corpora and in-domain unannotated texts.

We propose a unified model to learn from out-of-domain corpora and in-domain unannotated texts. The unified model contains two functions, one for cross-domain learning and another for semi-supervised learning. Our model can adjust learning rate in deep learning for every sentence.

For cross-domain learning, a lot of previous work focuses on domain similarity (Sun, Kashima, and Ueda 2013; Bhatt, Semwal, and Roy 2015; Bhatt, Sinha, and Roy 2016). Sun, Kashima, and Ueda (2013) propose a multitask learning method to automatically discover task relationships from real-world data. Their method can iteratively learn the task similarities via measuring the similarities of model weights of different tasks. They also show reasonable convergence properties by convergence analysis. Our cross-domain learning function can learn out-of-domain information based on domain similarity. We use similarity between out-of-domain sentence and in-domain corpus to adjust learning rate for every sentence in out-of-domain corpus.

For semi-supervised learning, a lot of previous work focused on confidence of prediction (Sarkar 2001; Watson and Briscoe 2007; Yu, Elkaref, and Bohnet 2015; Maeireizo, Litman, and Hwa 2004). We design a confidence based semi-supervised learning function, which can be used to learn in-domain unannotated information by self-training.

Our contributions in this work are as follow:

- We propose a unified model which can learn from out-of-domain corpora and in-domain unannotated texts.
- We propose a confidence based semi-supervised function which can learn in-domain unannotated texts by self-training.
- We propose a cross-domain function which can learn out-of-domain information based on domain similarity.

Background and Related Work

Our work focus on cross-domain and semi-supervised NER in Chinese social media with deep learning. We briefly review NER in Chinese social media, cross-domain learning and semi-supervised learning.

NER in Chinese Social Media

NER is a task to identify names in texts and to assign names with particular types (Sun et al. 2009; Sun 2014; Sun et al. 2014). As DEFT ERE Annotation Guidelines[1] shows, there are five entity types: person (PER), titles (TTL),

[1] Entities V1.7, Linguistic Data Consortium, 2014

organizations (ORG), geo-political entities (GPE) and locations (LOC). We consider PER, ORG, GPE and LOC. A mention is a single occurrence of a name (NAM), nominal phrase (NOM) or pronominal phrase (PRO) that refers to or describes a single entity. We consider NAM and NOM.[2] The main methods for NER treat it as a sequence tagging task.

It is difficult for NER in social media because of its informality and strong noise. There are many abbreviations and typos in social media texts. Furthermore, The Chinese language lacks explicit word boundaries, capitalization and other clues which are helpful for solving NER tasks in English. The difficulty and usefulness of NER in Chinese social Media attracts more and more attention. For example, Peng and Dredze (2015) explored several types of embeddings and proposed a joint training model for embeddings and NER; Peng and Dredze (2016a) used word segmentation representation to improve NER.

Cross-Domain Learning

Cross-domain learning methods need to make use of out-of-domain corpora to help improve in-domain results. There are several reasons why we need to pay attention to cross-domain tasks. First, it is hard to gain enough manually annotated texts for every domain, which costs a lot of time for annotation. Second, we may not know the domain of test data so we must consider domain adaption. However, in many NLP tasks, the performance will drop considerably when we test on a different domain, if we didn't design proper cross-domain learning methods. Fortunately, there is a lot pioneering work. For example, Lui and Baldwin (2011) chose to select cross-domain features to help domain adaption; Axelrod, He, and Gao (2011) chose to select pseudo in-domain data; Wen (2016) chose to train on multi-domain data. Sun, Kashima, and Ueda (2013) used Gaussian RBF and polynomial kernels to compute task similarity; Bhatt, Semwal, and Roy (2015) used cosine similarity measure to compute similarity for domain adaption.

Semi-Supervised Learning

In many NLP tasks, annotated data is quite limited but there are massive unannotated texts. Manual annotation will cost a lot of time, so it is important to explore methods to make use of unannotated data. There are many semi-supervised and unsupervised models. For example, self-training, co-training and tri-training are used to select most reliable data for training. Watson and Briscoe (2007) used confidence-based self-training to choose proper data; Sarkar (2001) used two models to choose confident unlabeled sentences; Yu, Elkaref, and Bohnet (2015) trained on source data and chose high confidence data from unlabeled data; Maeireizo, Litman, and Hwa (2004) used two classifiers to choose most confident instances; Kawahara and Uchimoto (2008) assessed reliability of predictions and selected most reliable predictions.

Proposal

We first build a bidirectional long short term memory neural network (BiLSTM) and combine transition probability to form structured output with max margin neural network (BiLSTM-MMNN). (Hochreiter and Schmidhuber 1997; Hammerton 2003; Chen et al. 2015; Taskar et al. 2005; Pei, Ge, and Chang 2014; Huang, Xu, and Yu 2015) Then, we propose a unified model for cross-domain and semi-supervised NER in Chinese social media. We explain our cross-domain learning function and semi-supervised learning function before our unified model.

BiLSTM-MMNN

We combine transition probability into BiLSTM with max margin neural network as our basic model.

Transition Probability We combine transition probability into BiLSTM with max margin neural network as our basic model. Max margin criterion concentrate directly on the robustness of decision boundary of a model, which will be easier to expand to our unified model.

We define a structured margin loss $\Delta(y, \bar{y})$ as Pei, Ge, and Chang(2014):

$$\Delta(y, \bar{y}) = \sum_{i=1}^{n} \kappa \mathbf{1}\{y_j \neq \bar{y_j}\} \tag{1}$$

where κ is the discount rate.

For a given instance x, our prediction will be the tag sequence with highest score:

$$y^* = \underset{\bar{y} \in Y(x)}{argmax}\, s(x, \bar{y}, \theta) \tag{2}$$

where $s(x, \bar{y}, \theta)$ is the score of tag sequence $\bar{y}$. The correct tag sequence y will be larger up to a margin to other possible tag sequences $\bar{y} \in Y(x)$:

$$s(x, y, \theta) \geq s(x, \bar{y}, \theta) + \Delta(y, \bar{y}) \tag{3}$$

To combine transition probability, our score function is as follow:

$$s(x, y, \theta) = \sum_{i=1}^{n} (A_{t_{i-1}t_i} + f_\Lambda(t_i|x)) \tag{4}$$

where $f_\Lambda(t_i|x)$ indicates the probability of tag t_i with parameters Λ, A indicates the matrix of transition probability and n is the length of sentence x.

In our model, $f_\Lambda(t_i|x)$ is computed as follow:

$$f_\Lambda(t_i|x) = -log(y_i[t_i]) \tag{5}$$

Character and Position Embeddings Word segmentation is important in Chinese text processing. Peng and Dredze (2015) explored three kinds of embeddings for NER in Chinese social media: word, character and character-positional embeddings. They showed that character-positional embeddings yielded best result. We choose character-positional embeddings in our models. For character-positional embeddings, it is based on character but also considers position of character in the word. It needs to segment word to get the character position in the word.

[2] Our work is close to mention detection but for simplicity, we use the term NER.

Cross-Domain Learning Function

It is hard to make use of out-of-domain corpora because of the difference between in-domain corpora and out-of-domain corpora. So we need to identify the similarity of out-of-domain sentences with in-domain corpus. For cross-domain learning, we train directly on both in-domain and out-of-domain data. But we use different learning rate for different out-of-domain sentences. The learning rate is adjusted by similarity function automatically. The learning rate for sentence x is computed as follow:

$$\alpha(x) = \alpha_0 * func(x, IN) \tag{6}$$

where α_0 is the fixed learning rate for in-domain sentences, $func(x, IN)$ indicates the similarity between sentence x and in-domain corpus IN, which is from 0 to 1.

In our model, we consider three different functions to compute the similarity.

Cross Entropy We consider cross entropy between sentence $x = W_1...W_N$ and in-domain n-gram language model LM_{IN}. The cross entropy similarity is computed as follow:

$$func(x, IN) = C\frac{1}{-\frac{1}{N}log_2(\prod_{i=1}^{N} P(W_i|W_{i-n+1}...W_{i-1}))} \tag{7}$$

where C is a real-valued constant for tuning the magnitude of similarity.

Gaussian RBF Kernel We consider Gaussian RBF kernel as follow:

$$func(x, IN) = \frac{1}{C}exp(-\frac{\|v_x - v_{IN}\|}{2\sigma^2}) \tag{8}$$

where C is a real-valued constant for tuning the magnitude of similarity, σ is used to control variance of the Gaussian RBF function, v_{IN} and v_x are vector representation for in-domain training data IN and sentence x. In our model, we first use word2vec (Mikolov and Dean 2013) to train on massive unannotated in-domain texts and gain embeddings for every character-positional. Sentence vector is the mean of character vectors in sentence. Corpus vector is the mean of sentence vectors in corpus.

Polynomial Kernel We consider polynomial kernel as follow:

$$func(x, IN) = \frac{1}{C}\frac{< v_x, v_{IN} >^d}{\|v_x\|^d . \|v_{IN}\|^d} \tag{9}$$

The definition of C, v_x and v_{IN} are the same as Gaussian RBF kernel. If $d = 1$, the normalized kernel has the form $\frac{1}{C}cos\theta$,where θ is the angle between v_x and v_{IN} in the Euclidean space, which is exactly the cos kernel.

Semi-Supervised Learning Function

Manual annotation costs a lot of time, so we need to try to make use of unannotated texts to help solve the task. There are many semi-supervised methods such as self-training, co-training, tri-training. The main purpose of these methods is to choose most confident prediction in the unannotated texts. We propose a semi-supervised learning function based on sentence confidence.

Our semi-supervised learning function is based on BiLSTM-MMNN which depends on the decision boundary. So sentence confidence of our model is based on the decision boundary. Our prediction is the tag sequence with highest score and the score need to be larger up to a margin to other possible tag sequences.

For sentence x, our prediction is the tag sequence with highest score as Equation 2:

$$y_{max}(x) = \underset{\bar{y} \in Y(x)}{argmax}\, s(x, \bar{y}, \theta)$$

we consider the tag sequence with the second highest score:

$$y_{2nd}(x) = \underset{\bar{y} \in Y(x)\ and\ \bar{y} \neq y_{max}}{argmax} s(x, \bar{y}, \theta)$$

Then our sentence confidence is defined as follow:

$$confid(x) = \frac{y_{max}(x) - y_{2nd}(x)}{y_{max}(x)} \tag{10}$$

In Equation 10, we can know if decision margin between the max and second sequence is larger, our prediction will be more confident.

Our semi-supervised learning function is dynamic because we compute confidence for sentences before every epoch. Because confidence is based on our model, confidence of sentences will be different in different epochs.

The learning rate $\alpha_t(x)$ for unannotated sentence x in epoch t is computed as follow:

$$\alpha^t(x) = \alpha_0^t * confid(x, t) \tag{11}$$

where α_0^t is the learning rate for in-domain sentences at epoch t, $confid(x, t)$ is the confidence of sentence x at epoch t.

Unified Model

In our unified model, learning rate $\alpha^t(x)$ for every sentence x at epoch t is computed as follow:

$$\alpha^t(x) = \alpha_0^t * weight(x, t) \tag{12}$$

where $weight(x, t)$ is used to adjust learning rate for sentence x. The definition of $weight(x, t)$ is as follow:

$$weight(x, t) = \begin{cases} 1.0 & \text{x is in-domain,} \\ func(x, IN) & \text{x is out-of-domain,} \\ confid(x, t) & \text{x is unannotated.} \end{cases}$$

where $func(x, IN)$ is the similarity between sentence x and in-domain corpus IN and $confid(x, t)$ is the confidence of sentence x at epoch t.

Experiments

To demonstrate the effectiveness our proposed model, we do some experiments on NER datasets. We will describe the details of datasets, settings and results in our experiments.

Models	Named Entity			Nominal Mention				
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Overall**	**OOV**
BiLSTM-MMNN	65.74	33.65	44.51	70.42	50.51	58.82	51.44	14.35
+ All Data Merge	43.58	45.02	44.29	28.81	17.17	21.52	32.27	**30.87**
Cross-Domain Learning (proposal)	52.94	**51.18**	52.05	71.63	51.01	59.59	55.70	**30.87**
Semi-Supervised Learning (proposal)	**68.42**	36.97	48.00	73.43	53.03	61.58	54.57	15.65
Unified Model (proposal)	61.68	48.82	**54.50**	**74.13**	**53.54**	**62.17**	**58.23**	28.70

Table 1: NER results for named and nominal mentions on test data. We can see that our cross-domain and semi-supervised learning improve NER. Our unified model outperforms previous work.

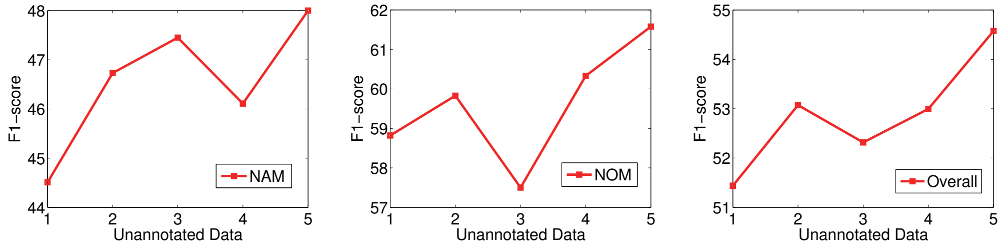


Figure 1: Comparing F1-scores with different amount unannotated data. We can see that our cross-domain model improves results with increased amount of unannotated data.

	Named Entity	**Nominal Mention**
Train set	957	898
Development set	153	226
Test set	211	198
Unlabeled Text	112,971,734 Weibo messages	

Table 2: Details of Weibo NER corpus.

Entity Type	**Train Set**	**Test Set**
Location	18522	3658
Organization	10261	2185
Person	9028	1864
Total	37811	7707

Table 3: Details of SIGHAN NER corpus.

Datasets

We use the same annotated corpus[3] as Peng and Dredze (2015; 2016a) for NER in Chinese social media. The corpus is composed of Sina Weibo[4] messages annotated for NER. The corpus contains PER, ORG, GPE and LOC for both named and nominal mention. For out-of-domain data, we use the MSR corpus of the sixth SIGHAN Workshop on Chinese language Processing. The SIGHAN corpus only contains LOC, PER and ORG three types for named mention. We also use the same unannotated texts as Peng and Dredze (2016a) from Sina Weibo service in China and texts are word segmented by a Chinese word segmentation system Jieba[5] as Peng and Dredze (2016a).

The details of Weibo NER corpus are shown in Table 2. For SIGHAN corpus, the details are shown in Table 3.

Baselines

We construct two baselines to compare with our proposed unified model. The first one is the BiLSTM-MMNN model trained and tested on in-domain corpus. As for the second one, we pre-train on out-of-domain data and then train on in-domain data. For simplicity, we use BiLSTM-MMNN, BiLSTM-MMNN + All Data Merge to denote the two baselines.

Settings

We pre-trained embeddings using word2vec (Mikolov and Dean 2013) with the skip-gram training model, without negative sampling and other default settings. Like Mao (2008), we use bigram features as follow:

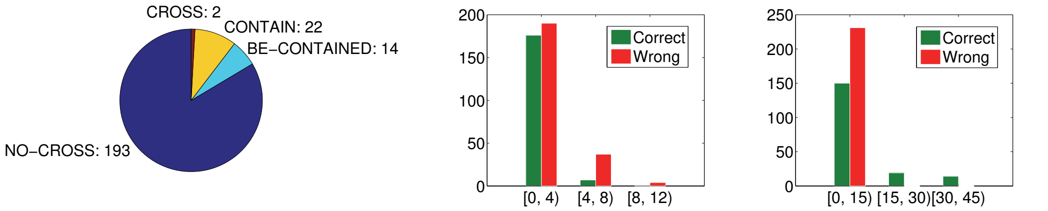
$$C_nC_{n+1}(n=-2,-1,0,1) \quad and \quad C_{-1}C_1$$

We use window approach (Collobert et al. 2011) to extract higher level features from word feature vectors. Our models are trained using stochastic gradient descent with $L2$ regularizer. As for parameters in our models, window size for word embedding is 5, word embedding dimension, feature embedding dimension and hidden vector dimension are all 100, discount κ in margin loss is 0.2, and the hyper parameter for the $L2$ is 0.000001. As for learning rate, the default learning rate α_0 is 0.1 with a decay rate 0.95. We set learning rate $\alpha_0 = 0.05$ in our unified model and $\alpha_0 = 0.003$ in + All Data Merge model. We train 10 epochs and choose the best prediction for test.

[3] We fix some annotating errors of the corpus.

[4] One of the most popular Chinese social media, which is similar to twitter in English.

[5] https://github.com/fxsjy/jieba.



(a) Error types of entities.

(b) Length of entities.

(c) Occurrence number in training data.

Figure 2: Results of basic model. Green bars denote correct predictions and red bars denote wrong predictions. We can know that our main error type is CROSS and basic model is not good at long entities or entities with few occurrences in training data.

	Named Entity	Nominal Mention
Original	**52.98**	33.06
Post process	52.05	**59.59**

Table 4: Difference between before and after post process. We can see that post process is helpful for nominal mention.

Results

Table 1 shows results in NER in terms of precision, recall, F1-score for NAM and NOM. We also consider micro F1-score and out-of-vocabulary entities (OOV) recall. We can know that out-of-domain data helps a lot in OOV recall.

Cros-domain Learning Function For cross-domain learning, we do experiments on the three similarity functions: cross entropy, Gaussian RBF kernel and Polynomial Kernel. For all similarity functions, we set the magnitude tuning constant $C = 1$. For cross entropy, we use trigram language model. We set $\sigma = 1$ for Gaussian RBF kernel. For polynomial kernel, we try different values of d and find that $d = 1$ worked well in our task. By comparing results of three similarity functions, we find that polynomial kernel with $d = 1$ get the best result. So we choose the polynomial kernel with $d = 1$ for our cross-domain learning function.

Because SIGHAN only contains named mentions, our model improves a lot in named mentions but compromises in nominal mentions. We use a post process to combine the result of cross-domain learning function and basic model BiLSTM-MMNN. The process will keep the prediction for nominal mentions of BiLSTM-MMNN and then adopt the prediction for named mentions of cross-domain learning function. The F1-scores of named and nominal mentions before (Original) and after post process are shown in Table 4.

Semi-Supervised Learning Function There are so many unannotated texts, so sentences in unannotated texts may be quite different from annotated sentences. We first select the sentences with highest cross-entropy similarity. To avoid that selected sentences almost have no entities, we use basic model to test first and choose equal proportion for sentences with / without entities[6].

	Named Entity	Nominal Mention
Cross-domain	52.98	33.06
Unified	**55.35**	**36.58**

Table 5: Comparing cross-domain learning and unified model. We can see that our cross-domain model outperforms basic model.

We experiment on our semi-supervised learning function with different amount unannotated sentences. We find that it is hard for us to make use of too many unannotated sentences.[7] We choose 1, 2, 3, 4, 5 times unannotated data of annotated data. The results with different amount unannotated data are shown in Figure 1, where NAM denotes named entity, NOM denotes nominal mention and Overall F1-score denotes micro F1-score of NAM and NOM.

Because we need to predict unannotated sentences before each epoch, we use pre-trained parameters of BiLSTM-MMNN to initialize parameters in our semi-supervised learning function.

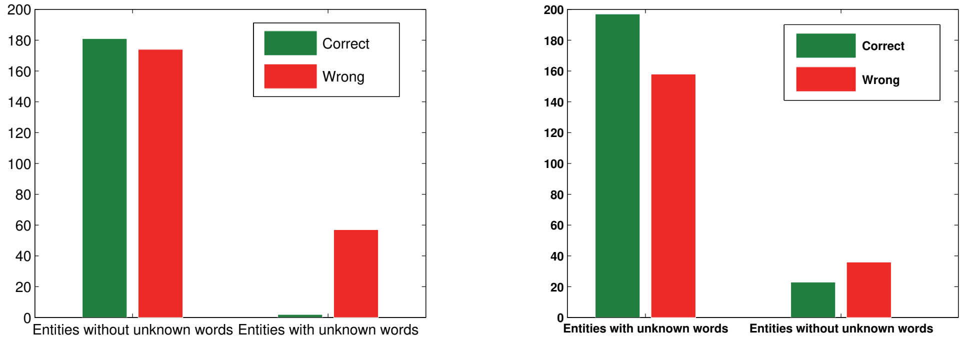
Unified Model Because our unified model use SIGHAN corpus, it also faces the same problem with cross-domain learning function. It also improves a lot in named mentions but compromises in nominal mention. We use the same post process to combine the result of unified model and semi-supervised learning function. Because unified model can make use of out-of-domain and unannotated data but semi-supervised learning function only requires unannotated data. We can build our unified model based on the prediction of semi-supervised learning function. To understand the advantage of our unified model, we compare the original results of cross-domain learning function and unified model as Table 5.

Error Analysis

Although our unified model outperforms previous work, we still need to know that result in Chinese social media is much lower than formal domains. For example, the state-of-art result of NER in SIGHAN is 92.81. The result in Weibo is much lower than it, so we need to do more analysis to help

[6] we found that the proportions of sentences with / without entities are similar in both Weibo and SIGHAN corpora.

[7] In our experiments, when unannotated data is 8 times bigger than annotated data, it will not be able to improve results.



(a) Unknown word rate of basic model.

(b) Unknown word rate of cross-domain learning.

Figure 3: Comparing basic and cross-domain model. We can see that our cross-domain model helps a lot for entities with unknown words.

understand why result in Weibo is so low. We also need to know why our model can outperform previous work.

We design six metrics to do error analysis as follows:

- Sentence length.
- Entity length.
- Five error types: CONTAIN[8], BE-CONTAINED[9], SPLIT[10], CROSS[11], NO-CROSS[12].
- Occurrence number in training data.
- Unknown word rate of sentence.
- Unknown word rate of entity.

BiLSTM-MMNN

By analyzing results of our basic model BiLSTM-MMNN, we find that entity length, occurrence number in training data, unknown word rate of entity have a great impact on prediction. We also find that NO-CROSS error type covers most of errors in our prediction. We show details of these analysis in Figure 2.

From Figure 2(a), we can find that no wrong predictions belong to SPLIT error type, which means that our predictions are continuous. The percentage of NO-CROSS errors is 83.55%. So it may be a good choice to focus on NO-CROSS errors. From Figure 2(b) and Figure 2(c), we can know that our basic model works not good when entity is long or entity appears little in training data.

Cross-Domain Learning Function

To compare results of basic model and cross-domain learning function, we find some improvement as follows:

- Cross-domain learning function causes a great decrease in NO-CROSS errors.
- Cross-domain learning function improves a lot for entities with few or no occurrences in training data.
- Cross-domain learning function helps a lot in predicting entities with unknown words.

From these improvement, we can know that out-of-domain can provide much more information about words and entities which are not in in-domain data. It is a good choice to use out-of-domain data to broaden knowledge of models. We compare results of basic model and cross-domain learning function in Figure 3.

Semi-Supervised Learning Function

By comparing results of basic model and semi-supervised learning function, we find that there is no big improvement in a special aspect but semi-supervised enhance overall results. For semi-supervised learning function, we use self-training method to assign a confidence for every unannotated sentence. Unannotated sentences are used to strengthen our training on annotated data. So it helps almost every aspect but the improvement is not big.

Unified Model

Our unified model combines cross-domain learning and semi-supervised learning which are better than two learning functions. But the results are still quite low, when comparing to results in SIGHAN.By analyzing results of our unified model, we still need to focus on following points:

- NO-CROSS errors.
- Named or nominal mentions with few or no occurrences in training data.
- Entities with high unknown word rate or entities in sentence with high unknown word rate.
- Long entities or entities in long sentence.

Conclusions

We propose a unified model for NER in Chinese social media. The model can learn from out-of-domain corpora and

[8] Gold one contains our prediction

[9] Gold one is contained by our prediction

[10] There are gaps in our prediction

[11] Gold one cross our prediction

[12] There are no common words between gold one and our prediction

in-domain unannotated texts. In our experiments, our unified model outperforms previous work. Furthermore, our targeted and detailed error analysis not only helps us understand the advantage of our model but also points out the aspects we need to pay more attention to.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (No. 61300063, No. 61673028), and National High Technology Research and Development Program of China (863 Program, No. 2015AA015404). Xu Sun is the corresponding author of this paper.

References

Axelrod, A.; He, X.; and Gao, J. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of EMNLP*, 355–362.

Bhatt, H. S.; Semwal, D.; and Roy, S. 2015. An iterative similarity based adaptation technique for cross domain text classification. *CoNLL 2015* 52.

Bhatt, H. S.; Sinha, M.; and Roy, S. 2016. Cross-domain text classification with multiple domains and disparate label sets. In *ACL (1)*.

Chen, X.; Qiu, X.; Zhu, C.; Liu, P.; and Huang, X. 2015. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of EMNLP*, 1385–1394.

Cherry, C., and Guo, H. 2015. The unreasonable effectiveness of word representations for twitter named entity recognition. In *HLT-NAACL*, 735–745.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12(Aug):2493–2537.

Habib, M. B., and van Keulen, M. 2015. Need4tweet: a twitterbot for tweets named entity extraction and disambiguation.

Hammerton, J. 2003. Named entity recognition with long short-term memory. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 172–175.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Kawahara, D., and Uchimoto, K. 2008. Learning reliability of parses for domain adaptation of dependency parsing. In *IJCNLP*, volume 8.

Li, C., and Liu, Y. 2015. Improving named entity recognition in tweets via detecting non-standard words. In *ACL (1)*, 929–938.

Lui, M., and Baldwin, T. 2011. Cross-domain feature selection for language identification. In *In Proceedings of 5th International Joint Conference on Natural Language Processing*.

Maeireizo, B.; Litman, D.; and Hwa, R. 2004. Co-training for predicting emotions with spoken dialogue data. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 28.

Mao, X.; Dong, Y.; He, S.; Bao, S.; and Wang, H. 2008. Chinese word segmentation and named entity recognition based on conditional random fields. In *IJCNLP*, 90–93.

Mikolov, T., and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Pei, W.; Ge, T.; and Chang, B. 2014. Max-margin tensor neural network for chinese word segmentation. In *ACL (1)*, 293–303.

Peng, N., and Dredze, M. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of EMNLP*, 548–554.

Peng, N., and Dredze, M. 2016a. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 149–155.

Peng, N., and Dredze, M. 2016b. Multi-task multi-domain representation learning for sequence tagging. *arXiv preprint arXiv:1608.02689*.

Sarkar, A. 2001. Applying co-training methods to statistical parsing. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, 1–8. Association for Computational Linguistics.

Sun, X.; Matsuzaki, T.; Okanohara, D.; and Tsujii, J. 2009. Latent variable perceptron algorithm for structured classification. In *IJCAI 2009*, 1236–1242.

Sun, X.; Li, W.; Wang, H.; and Lu, Q. 2014. Feature-frequency-adaptive on-line training for fast and accurate natural language processing. *Computational Linguistics* 40(3):563–586.

Sun, X.; Kashima, H.; and Ueda, N. 2013. Large-scale personalized human activity recognition using online multitask learning. *IEEE Transactions on Knowledge and Data Engineering* 25(11):2551–2563.

Sun, X. 2014. Structure regularization for structured prediction. In *Advances in Neural Information Processing Systems 27*, 2402–2410.

Taskar, B.; Chatalbashev, V.; Koller, D.; and Guestrin, C. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, 896–903.

Watson, R., and Briscoe, T. 2007. Adapting the rasp system for the conll07 domain-adaptation task. In *EMNLP-CoNLL*, 1170–1174. Citeseer.

Wen, T.-H.; Gasic, M.; Mrksic, N.; Rojas-Barahona, L. M.; Su, P.-H.; Vandyke, D.; and Young, S. 2016. Multi-domain neural network language generation for spoken dialogue systems. *arXiv preprint arXiv:1603.01232*.

Yu, J.; Elkaref, M.; and Bohnet, B. 2015. Domain adaptation for dependency parsing via self-training. *IWPT 2015* 1.